W484–W490 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 7 June 2012*
doi:10.1093/nar/gks458

IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks

Aaron K. Wong[1], Christopher Y. Park[1], Casey S. Greene[2], Lars A. Bongo[3], Yuanfang Guan[2,4,5] and Olga G. Troyanskaya[1,2,*]

[1]Department of Computer Science, [2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA, [3]Department of Computer Science, University of Tromsø, N-9037 Tromsø, Norway, [4]Department of Computational Medicine and Bioinformatics and [5]Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA

Received January 31, 2012; Revised April 20, 2012; Accepted April 30, 2012

ABSTRACT

Integrative multi-species prediction (IMP) is an interactive web server that enables molecular biologists to interpret experimental results and to generate hypotheses in the context of a large cross-organism compendium of functional predictions and networks. The system provides a framework for biologists to analyze their candidate gene sets in the context of functional networks, as they expand or focus these sets by mining functional relationships predicted from integrated high-throughput data. IMP integrates prior knowledge and data collections from multiple organisms in its analyses. Through flexible and interactive visualizations, researchers can compare functional contexts and interpret the behavior of their gene sets across organisms. Additionally, IMP identifies homologs with conserved functional roles for knowledge transfer, allowing for accurate function predictions even for biological processes that have very few experimental annotations in a given organism. IMP currently supports seven organisms (*Homo sapiens*, *Mus musculus*, *Rattus novegicus*, *Drosophila melanogaster*, *Danio rerio*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*), does not require any registration or installation and is freely available for use at http://imp.princeton.edu.

INTRODUCTION

As high-throughput experiments become increasingly common, biologists face substantial challenges effectively leveraging genome-scale data from diverse organisms to inform new hypotheses. Experimental data coverage for an organism can be sparse, and prior functional knowledge (i.e. low-throughput experiments validating a gene's function) can be notably limited. These impediments affect the breadth and accuracy of bioinformatic methods (e.g. machine-learning algorithms) that apply prior knowledge in learning novel biology. As a consequence, the applicability of these methods is often limited to biological processes and pathways that are already well characterized for an organism.

For example, a common challenge for biological researchers is interpreting the results of a genome-wide experiment (e.g. a list of candidate genes from a microarray experiment) and generating hypotheses for experimental follow-up. There are several effective resources, some network based, for researchers to analyze their gene sets (1–6). These resources cover a wide range of organisms and address different needs of biologists by applying a variety of methods: from pathway enrichment analysis of a gene list to machine learning algorithms that predict a gene's function. All these resources' methods require known examples (i.e. pathways with at least a few annotated genes) in an organism. Consequently, the effectiveness of these applications is constrained by the extent of prior knowledge and available experimental data in the queried organism.

Other resources address the problem of disparate data coverage among organisms by focusing on methods to transfer high-throughput data (e.g. microarray and physical interaction experiments) between organisms (7,8). However, these efforts are limited to learning gene association networks, and none of them solve the problem of making accurate functional predictions and associations for biological processes that have not been well

*To whom correspondence should be addressed. Tel: +1 609 258 1749; Fax: +1 609 258 1771; Email: ogt@genomics.princeton.edu

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

studied in a given organism. For example, most of the discovered genes involved in neuromuscular process have been in mouse [65 known genes according to gene ontology (9)]. Relatively few genes are definitively known in mammalian systems outside of this model organism. Consequently, many existing methods will not be able to predict genes to that biological process in rat (where only one such gene is experimentally annotated), and a biologist using a rat model system with existing resources will not be able to leverage the known biology in mouse. Biologists need a technology that allows for the systematic application of prior functional knowledge from other organisms to their organism of study, at multiple points in an analytic workflow: from interpreting experimental results to generating hypotheses for functional assays. Integrative multi-species prediction (IMP) is an interactive web server designed to meet this burgeoning need.

IMP is an exploratory tool that, in addition to providing a high-quality interface for functional interrogation, solves several specific challenges encountered by biologists that benefit from integrating cross-organism biological knowledge. First, although biologists can interpret their experimental results in the context of functional networks, other servers do not allow them to adequately accomplish this task due to their limited workflow support and the incomplete prior knowledge in an organism. With IMP, biologists can save their custom genes sets and overlay their genes on functional networks, expanding or focusing their gene list by mining functional relationships within the networks. IMP can integrate cross-organism knowledge with a method that goes beyond the standard Basic Local Alignment Search Tool (10) search by identifying enriched biological processes among the genes, using gene-pathway annotations from the queried organism and annotations from other organisms mapped by functional analogy (11). In this way, pathways that are better characterized in a different organism will be included in an enrichment analysis, facilitating biological connections that would otherwise be hindered by limited functional knowledge. Moreover, no existing server provides a way to interactively examine putative functions and gene–gene interactions in functional networks across organisms. With IMP, biologists can compare functional contexts and interpret the behavior of their gene sets across organisms using flexible and interactive visualizations.

Finally, the results from a genome-wide experiment can elucidate a biological question but are often inconclusive and require experimental follow-up. Computational predictions of gene function can guide subsequent experiments. However, accurate assignments have previously been limited to pathways and processes that are already well characterized in an organism, as such information is necessary for training examples. Transferring functional annotations between homologous genes is a common method to improve coverage for a studied pathway and to generate hypotheses for functional assays, but high-quality transfers have historically been limited to smaller scale, manual curation efforts (12). IMP systematically identifies functionally similar homologs, using state-of-the-art homolog identification methods that use genomic data compendia (11) to transfer pathway annotations between organisms for learning. This allows for accurate gene-process predictions, even for processes that have few experimental annotations in an organism.

IMP SYSTEM DETAILS

An analysis in IMP begins with a biological process, gene or set of genes of interest. The biological questions that IMP can help answer depend on the provided input. For example, a researcher may have a group of co-expressed genes from an messenger ribonucleic acid experiment and want to determine the common biological functions and pathways among the genes. The researcher may also want to compare the behavior of this gene set in different organisms and identify additional genes predicted to be interacting with the input set. Alternatively, a biologist may be interested in a specific pathway and seek additional candidate genes that may be involved in or interacting with this pathway to investigate. Using IMP, biologists can answer these questions with the functional knowledge acquired in the organism of interest and potentially better-studied organisms.

IMP currently supports seven organisms, with plans to add more in future updates. The functional networks are constructed using previously described methods (4,5,13) and integrate genomic data from an array of public data sources (9,14–17). These data can cover diverse tissues and developmental time points. Our integration method summarizes them into a global picture of gene–gene relationships. The complete list of data sources is accessible directly on the web server (http://imp.princeton.edu/networks/data/).

Although functional networks of IMP are constructed using organism-specific data, IMP leverages biological knowledge from multiple organisms in several functional analyses. To accomplish this, IMP transfers functional knowledge (experimental gene ontology (GO) biological process annotations) using our previously developed method to identify appropriate homologs for annotation transfer (11). This method from Chikina and Troyanskaya (11) extends beyond simple annotation transfer by sequence similarity. Instead, a network-based similarity score is used to identify all homologs [defined as genes separated by a speciation or duplication event in a TreeFam family (18)] with similar functional profiles. In effect, a more specific relationship is identified for annotation transfer than when using sequence alone—functional knowledge is transferred between genes with a shared evolutionary history that also exhibit similar biological function.

ANALYZING CUSTOM GENE SETS

With IMP, biologists create custom gene sets for any organism, assign them informative labels and colors and submit genes for analysis (Figure 1). These gene sets can be saved on the user's local machine by cookies and persist

as long as cookies are kept (usually one year by default), or they can be saved to a server and shared with collaborators. In the analysis, a graphical search of an organism's gene network is performed on the gene set to retrieve predicted functional neighbors (those likely to participate in the same pathway). Small gene sets (1–10 genes), which have been reported to be the majority of user inputs to other web servers (2), benefit particularly from this network-based analysis. A small and statistically underpowered gene set can be expanded with functionally similar genes to improve biological interpretation and meet significance cutoffs for pathway enrichment.

IMP presents these results as highly interactive networks: users can adjust layout by moving genes, query evidence supporting a functional relationship by hovering over an edge and modify graphical options to customize the display. Users can control the number of genes displayed by adjusting the confidence cutoff for the returned functional relationships or by filtering based on connectivity in the returned network. An enrichment analysis updates in real time as the network specificity is narrowed or broadened. The analysis identifies overrepresented pathways among the displayed genes using annotations from numerous public databases (9,17,19,20). Statistical significance is calculated using the hypergeometric distribution, and multiple test correction is performed using false-discovery rates (21).

The enrichment analysis can incorporate pathway annotations transferred from other organisms. The appropriate homologs of the network genes are identified, and any corresponding pathway annotations can be included. Thus, pathways that are not well characterized in a biologist's organism of interest, but better studied in another organism, will be included in the results.

Finally, any gene set created by the user persists throughout IMP and is automatically overlaid on any relevant network presented to the user. Any gene from a custom set is rendered with the set's user-defined color and labeled in the network legend. In this way, IMP serves

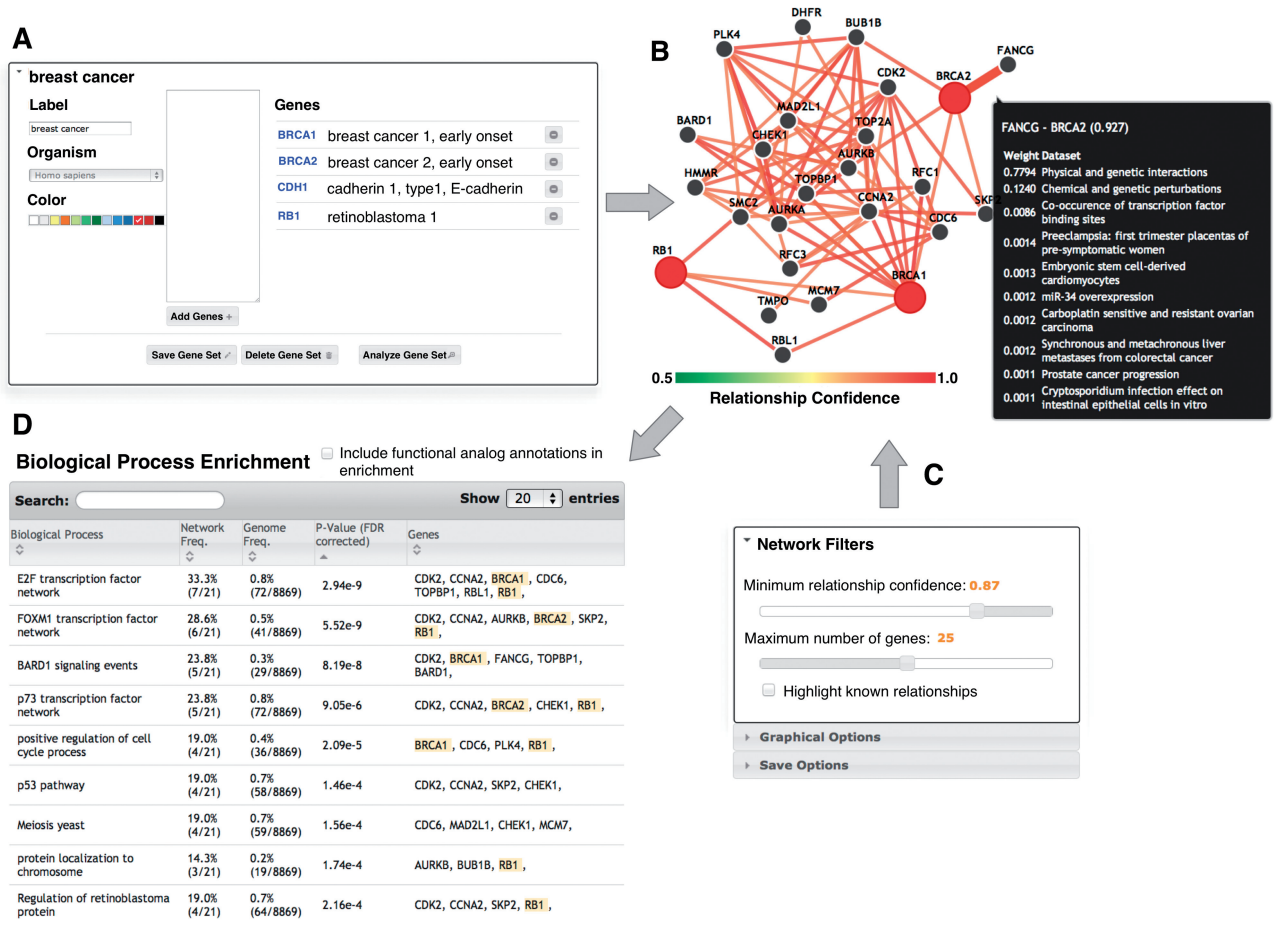

Figure 1. The workflow for creating and analyzing a custom gene set in IMP. (**A**) The interface for creating a gene set, where users can select an organism, add their genes and assign a label and informative color. This can be reached by following the 'My Gene Sets' link at the top right of every page. (**B**) The returned sub-network for the submitted gene set (large nodes), with the nodes colored using the user-assigned color for the custom set (red). Edge colors correspond to the confidence of a functional relationship between genes. Users can hover over any edge to examine the top data sets contributing to that score. (**C**) The added genes in the displayed network are functionally related to the queried genes at a confidence cutoff controlled by the user. (**D**) A biological process enrichment calculation is updated in real time as genes are added or removed from the network display. The user can choose to include annotations transferred from other organisms in the enrichment calculation.

as an exploratory tool that is also anchored by a user's experimental results or genes of interest.

CROSS-ORGANISM NETWORK ANALYSIS

The accuracy and coverage of gene association networks depend largely on the extent of publicly available experimental data. Because some organisms are developed to elucidate certain biological processes, data coverage for a studied pathway can vary considerably by organism. Using IMP, biologists can compare functional contexts of interest, whether a user submitted gene set or a previously defined pathway in IMP, across multiple organisms. They can visually assess whether the predicted functional relationships for their gene set are supported in other organisms and can discover what data support connections in each organism.

The cross-organism network layout is shown in Figure 2. The relevant process network for a user-submitted gene set is visually aligned against the networks of other organisms. The alignment uses a coordinated layout where genes are mapped by functional analogy (11) and user interactions with the queried network are simultaneously updated in the other organisms' networks. This interactive view of predicted pathways across organisms can help answer broad and important biological questions. In particular, this helps biologists address a key challenge in the study of human disease: identifying the best model system for a given disease, with the most appropriate ortholog for a disease gene of interest.

For example, Figure 2 shows the pathway for double-strand break repair, a critical process that, when defective, can lead to genome rearrangements that ultimately lead to cancerous cells [reviewed in (22)]. This process is highly conserved across evolutionarily distant organisms, which is reflected in the network comparison. Data coverage is high across the three organisms (human, mouse and yeast), as evidenced by the high confidence edges between most of the genes in the pathway. Moreover, it is apparent that many genes (e.g. RAD51, POLA1 and RPA2) have conserved functional relationships and would be potentially good candidates for study in mouse or yeast.

KNOWLEDGE-TRANSFERRED FUNCTION PREDICTIONS

A biologist can also use IMP to generate hypotheses for follow-up experiments. Users can query IMP by gene or GO biological process to retrieve gene-pathway predictions. IMP applies a previously developed and validated method (23), which uses a functional network as input to a support vector machine (SVM), to classify genes to biological processes. The predictions in IMP leverage the functional knowledge from multiple organisms by using the transferred and directly assigned annotations in an organism as positive examples in the SVM (Supplemental Figure S1). An SVM is trained with 5-fold cross-validation for each GO biological process term using this expanded set of annotations. Thus, IMP can accurately assign genes to processes that have few experimental annotations in a queried organism.

Figure 3A shows the result page for querying BRCA2 in human. The relevant local network of predicted functional relationships and putative biological processes are returned to the user. All the networks visualized in IMP can be exported as a high-quality figure (SVG format) for inclusion in publications. IMP result pages use interactive elements and advanced visualizations. These features make information accessible that would otherwise be cumbersome to present. For example, a user can click on the edge connecting BRCA1 and BRCA2 in the network to bring up a detailed view of the top data sets contributing to the predicted functional relationship. Additionally, users can search the list of predicted processes and sort or filter by process specificity. Researchers can perform all these tasks without leaving the result page through the dynamic, interactive elements.

In addition to queries by gene, a scientist can explore a biological process of interest (Figure 3B). The result page for a biological process query contains a list of genes predicted to participate in the process and the functional network of genes already annotated to the process. Users can click on a gene's description to update the displayed network with functional relationships between the selected gene and the genes used as positive examples in the classifier. In this way, a researcher can visually assess the relationships that support the prediction of the gene to the biological process.

The predictions in IMP have been validated using a conservative 1-year holdout evaluation: all biological process annotations available on June 2010 are used for training the SVM classifiers and annotations made in the subsequent year (through June 2011) are held out for evaluation. We assess performance with standard machine learning metrics (Supplemental Figure S2), and our performance is competitive among state-of-the-art function prediction methods (Supplemental Figure S3), even though many of these GO terms have too few annotations to make predictions without our annotation transfer method. Regularly updated plots of prediction performance are available directly on the web server.

CASE STUDY: THE FUNCTIONAL ROLE OF EVE1 IN *DANIO RERIO*

An example workflow illustrates some (though not all) of the capabilities of IMP. eve1 is a *Danio rerio* transcription factor implicated in a conserved role in body patterning in a variety of species (24). However, little is known about the biological mechanism of eve1 beyond its global role in patterning. A researcher may be interested in identifying the specific processes that eve1 participates in.

To generate testable hypotheses for the functional role of eve1, a biologist can query IMP for the gene. They can search IMP using a variety of identifiers for eve1 (Entrez, Ensembl and Zfin), in addition to its standard name. The result page for this query contains a list of biological processes predicted for eve1 and a network with genes

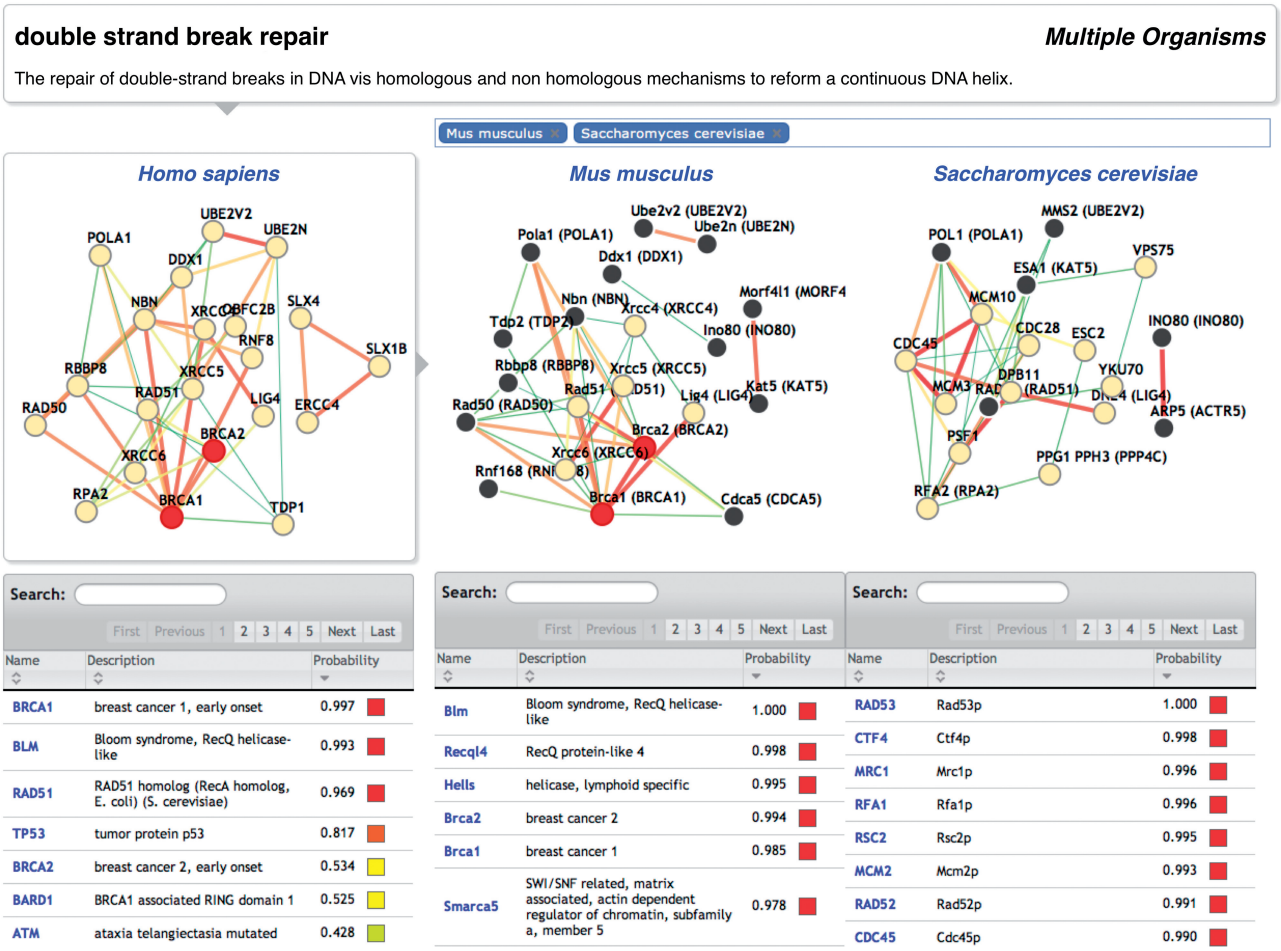

Figure 2. The cross-organism network view in IMP. The functional networks for the double-strand break repair pathway are visually aligned, and any interactions in the queried network (human) are simultaneously updated in the other networks (mouse and yeast). Yellow nodes indicate genes that are annotated to the double-strand break repair pathway in the respective organism. Gray nodes are genes with a homolog in the query organism.

predicted to be functionally related. Consistent with prior functional knowledge, the top prediction for eve1 is pattern specification process (GO:0007389), which reflects our understanding of eve1 in patterning. To identify a more specific functional role of eve1, the biologist can filter the prediction list by process size directly from the result page. The biological process of anterior–posterior pattern formation (122 genes) is both more specific than its grandparent in the GO hierarchy, pattern specification process (299 genes), and predicted with high confidence (85% probability).

In a recent publication, with experimental results not included in IMP, authors show that eve1 is an important regulator in posteriorisation and neural induction through a series of loss and gain of function experiments (25). Thus, in this case, IMP predicts a functional role (anterior–posterior pattern formation) for eve1 that was independently confirmed. Additionally, although the role of eve1 in neural induction was not directly predicted, a higher level but related process was predicted with high confidence (central nervous system development with 78% probability).

SUMMARY

IMP is a highly interactive and flexible web server that serves as an intuitive and accessible resource for functional interrogation. It addresses the current limitations of other web resources by integrating the prior knowledge and current experimental data from other organisms in its analyses. Its functional networks and predictions are generated using an automated analysis pipeline and will be updated semi-annually with additional organisms and data sets. The user interface is under active development, and features are constantly added based on the feedback. By automating the transfer of functional annotations between organisms and integrating this expanded knowledge in its analysis tools, IMP provides a unique suite of resources for biological researchers.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–3 and Supplementary Reference [26].

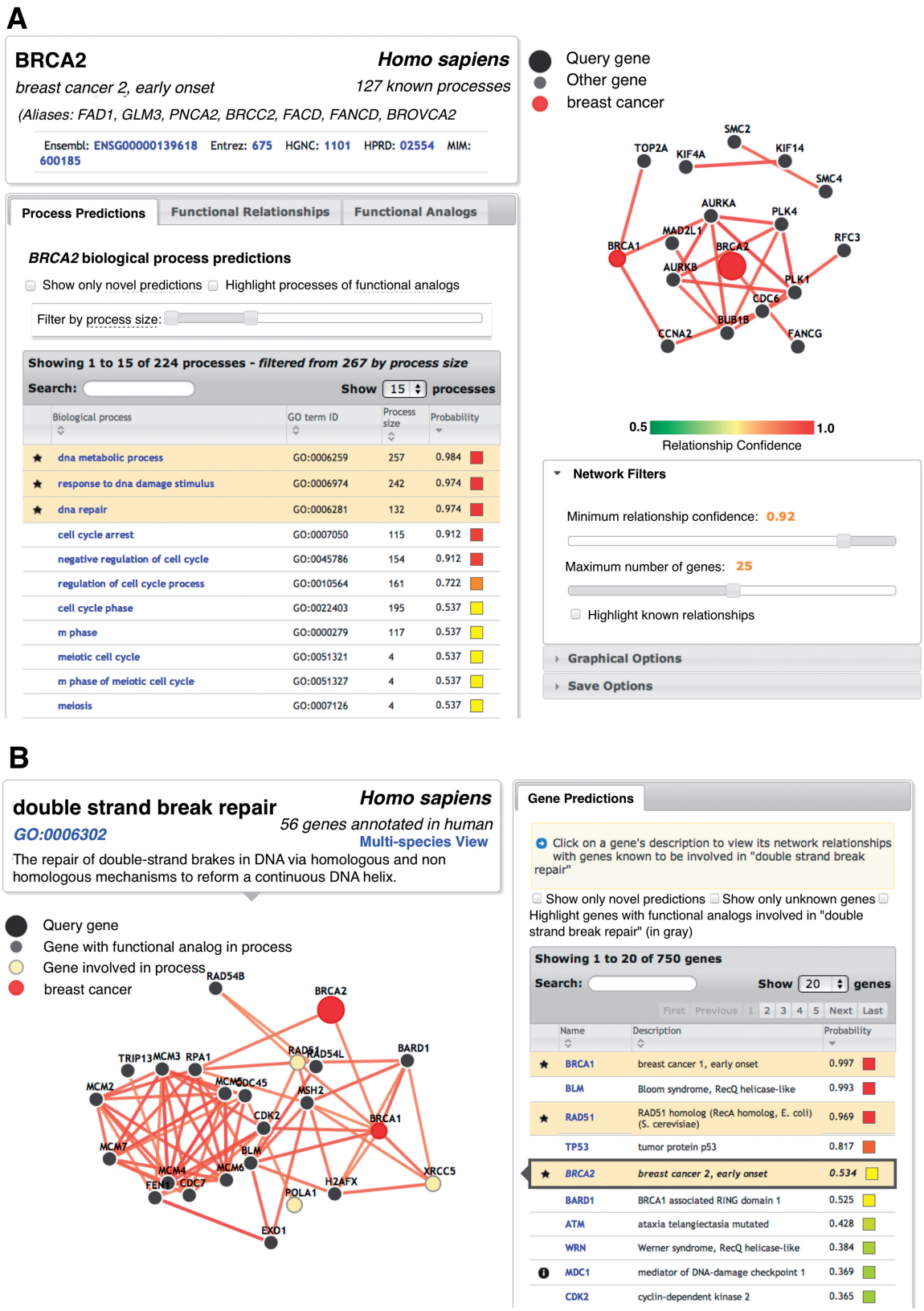

Figure 3. The result pages for a gene and biological process query. (**A**) A gene query returns the relevant local network and a list of biological processes predicted for the gene. The predicted processes can be searched by name or filtered by specificity (process size). (**B**) A biological process query returns a list of predicted genes likely to participate in the process. Clicking on a gene description will update the network on the left with relationships between the selected gene (large red node) and genes already known to be involved in the process (yellow nodes).

ACKNOWLEDGEMENTS

The authors acknowledge John Wiggins for his technical support, Kara Dolinski for helpful discussions and Josh Eidelson for critical comments on the manuscript.

FUNDING

National Science Foundation (NSF) CAREER [award DBI-0546275]; National Institutes of Health (NIH) [R01 GM071966, R01 HG005998 and T32 HG003284]; National Institute of General Medical Sciences (NIGMS) Center of Excellence [P50 GM071508]. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler: a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*., **35**, W193–W200.
2. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*., **4**, 44–57.
3. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al*. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*., **38**, W214–W220.
4. Myers,C.L., Robson,D., Wible,A., Hibbs,M.A., Chiriac,C., Theesfeld,C.L., Dolinski,K. and Troyanskaya,O.G. (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol*., **6**, R114.
5. Guan,Y., Myers,C.L., Lu,R., Lemischka,I.R., Bult,C.J. and Troyanskaya,O.G. (2008) A genomewide functional network for the laboratory mouse. *PLoS Comput. Biol*., **4**, e1000165.
6. Kao,H.-L. and Gunsalus,K.C. (2008) Browsing multidimensional molecular networks with the generic network browser (N-Browse). *Curr. Protoc. Bioinformatics*, Chapter 9, Unit 9.11.
7. Mering,von,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.
8. Alexeyenko,A., Schmitt,T., Tjärnberg,A., Guala,D., Frings,O. and Sonnhammer,E.L.L. (2012) Comparative interactomics with Funcoup 2.0. *Nucleic Acids Res*., **40**, D821–D828.
9. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.
10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
11. Chikina,M.D. and Troyanskaya,O.G. (2011) Accurate quantification of functional analogy among close homologs. *PLoS Comput. Biol*., **7**, e1001074.
12. Eisen,J.A., Sweder,K.S. and Hanawalt,P.C. (1995) Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res*., **23**, 2715–2723.
13. Huttenhower,C., Haley,E.M., Hibbs,M.A., Dumeaux,V., Barrett,D.R., Coller,H.A. and Troyanskaya,O.G. (2009) Exploring the human genome with functional maps. *Genome Res*., **19**, 1093–1106.
14. Stark,C., Breitkreutz,B.-J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al*. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*., **39**, D698–D704.
15. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al*. (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*., **40**, D857–D861.
16. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al*. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res*., **40**, D841–D846.
17. Caspi,R., Altman,T., Dreher,K., Fulcher,C.A., Subhraveti,P., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al*. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*., **40**, D742–D753.
18. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Hériché,J.-K., Osmotherly,L., Li,R., Liu,T., Zhang,Z., Bolund,L. *et al*. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*., **34**, D572–D580.
19. Kotera,M., Hirakawa,M., Tokimatsu,T., Goto,S. and Kanehisa,M. (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol*., **802**, 19–39.
20. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al*. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*., **37**, D619–D622.
21. Benjamini,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
22. Khanna,K.K. and Jackson,S.P. (2001) DNA double-strand breaks: signaling, repair and the cancer connection. *Nat. Genet*., **27**, 247–254.
23. Guan,Y., Ackert-Bicknell,C.L., Kell,B., Troyanskaya,O.G. and Hibbs,M.A. (2010) Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput. Biol*., **6**, e1000991.
24. Ahringer,J. (1996) Posterior patterning by the *Caenorhabditis elegans* even-skipped homolog vab-7. *Genes Dev*., **10**, 1120–1130.
25. Cruz,C., Maegawa,S., Weinberg,E.S., Wilson,S.W., Dawid,I.B. and Kudoh,T. (2010) Induction and patterning of trunk and tail neural ectoderm by the homeobox gene eve1 in zebrafish embryos. *Proc. Natl Acad. Sci. USA*, **107**, 3564–3569.
26. Peña-Castillo,L., Tasan,M., Myers,C.L., Lee,H., Joshi,T., Zhang,C., Guan,Y., Leone,M., Pagnani,A., Kim,W.K. *et al*. (2008) A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol*., **9(Suppl. 1)**, S2.